Supplementary Material

Table 1. Breakdown of tasks, longitudinality size of each dataset.

Task	Time Points			Counts (pCR/NpCR)	
	Pre-NAT	In-NAT	Post-NAT	Subjects	Scans
Generation	✓	✓		94(28/66)	1056(248/808)
	✓		✓	340(125/215)	3400(1024/2256)
pCR evaluation	✓			340(125/215)	3400(1024/2256)
	✓	✓(gI_2)			
	✓		✓		
	✓	✓(I_2)		94(28/66)	1056(248/808)

The pCR evaluation experiment using $I2$ is presented in supplementary S.Tab.2, with the remaining experiments based on the same experimental set shown in Tab.1 and Tab.2 of the main paper.

Table 2. pCR prediction performance, using generated in-NAT mammograms vs. real-world in-NAT mammograms (GT) based on the same experimental set. Each P-value is calculated on AUC by comparing it with the GT (the last column).

Methods	Sensitivity	Specificity	PPV	NPV	AUC	P-value
Diffusion-based model	0.476 [0.383,0.539]	0.762 [0.683,0.865]	0.691 [0.642,0.887]	0.602 [0.527,0.678]	0.651 [0.575,0.729]	1.831e-02
VAE-based model	0.474 [0.382,0.537]	0.747 [0.669,0.842]	0.634 [0.581,0.826]	0.541 [0.474,0.628]	0.616 [0.539,0.684]	8.021e-03
Ours	0.685 [0.593,0.738]	0.755 [0.678,0.845]	0.813 [0.764,0.890]	0.686 [0.609,0.750]	0.740 [0.668,0.811]	8.839e-01
GT	0.692 [0.605,0.752]	0.988 [0.866,0.997]	0.984 [0.878,1.000]	0.765 [0.685,0.837]	0.808 [0.727,0.879]	–

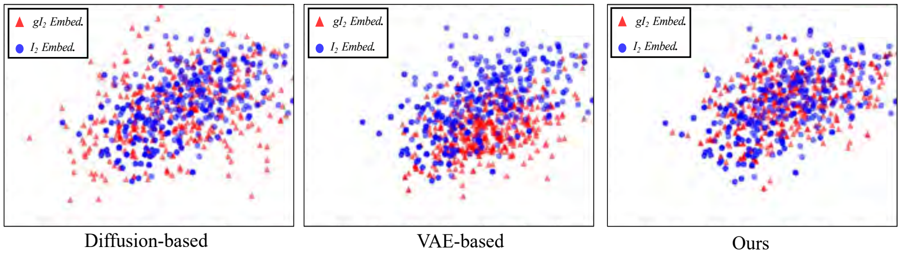

Fig. 1. t-SNE visualizations of generated ($gI2$) and real-world in-NAT mammogram ($I2$) representations. Each mammogram representation, with a shape of 1 × 2048, is embedded using a frozen ImageNet pre-trained ResNet-50.